Projet 2

Analysez des données de systèmes éducatifs

18 Juillet 2021

MBIADOU SALEU G. Raïssa

PLAN

01 **Rappel de la problématique et présentation du jeu de données**

Contexte, problématique, objectifs et description du dataset

02 **Analyse pré-exploratoire**

Environnement de développement, Analyse de la qualité, Sélection des données pertinentes, Scoring et Visualisation des résultats

03 **Conclusion et perspectives**

Suggestions de réponses aux questions métier et perspectives d'amélioration

01

PROBLÉMATIQUE ET JEU DE DONNÉES

Contexte, problématique, objectifs
et description du dataset

Contexte

- academy : start-up de la EdTech
- Métier: propose des **contenus de formation en ligne**
- Cible: public de niveau **lycée** et **université**
- Projet d'**expansion à l'international**
- Identification du **potentiel de nouveaux marchés** (en explorant un jeu de données de la Banque Mondiale sur le thème de l'éducation)

QUESTIONS

- Quels sont les pays avec un fort potentiel de clients pour les services de l'entreprise ?
- Pour chacun de ces pays, quelle sera l'évolution de ce potentiel de clients ?
- Dans quels pays l'entreprise doit-elle opérer en priorité ?

Objectifs de la mision

Validation de la qualité du jeu de Données

Analyser la qualité du jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)

Description du contenu du jeu de données

Décrire les informations contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)

Selection des données pertinentes

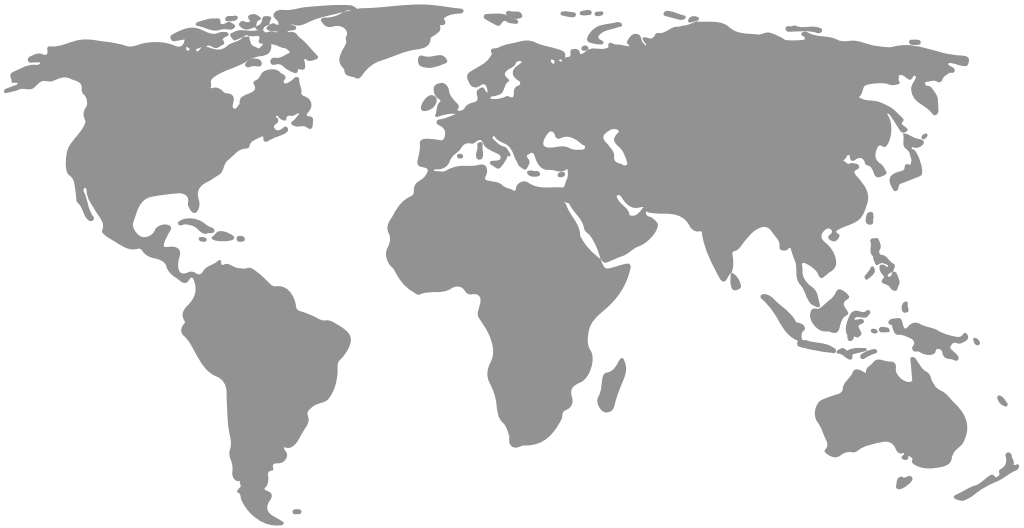
Sélectionner le colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise

Analyse pré-exploratoire

Déterminer des ordres de grandeurs des indicateurs statistiques classiques pour les différentes zones géographiques et pays du monde

Source du jeu de données

Collectées par l'organisme **"EdStats All Indicator Query"** de la **Banque mondiale** sur des sujets liés à l'éducation



Indicateurs

3665 indicateurs internationaux

Sujets

Indicateurs répartis en 37 sujets

Régions

7 régions du globe

Pays

242 pays et zones

1970-2100

Description

FICHIER	#LIGNES	#COLONNES	DESCRIPTION
EdStatsData.csv	886930	70	Valeurs des indicateurs pour tous les pays (1970-2100)
EdStatsSeries.csv	3665	21	Infos sur les indicateurs regroupés par sujets (topics)
EdStatsCountry.csv	241	32	Infos globales sur l'économie de chaque pays/zone du monde
EdStatsCounty-Series.csv	613	4	Description des sources des indicateurs
EdStatsFootnote.csv	643638	5	Infos supplémentaires indicateurs / pays / année

02

ANALYSE PRÉ-EXPLORATOIRE

Environnement de développement, Analyse de la qualité, Sélection des données pertinentes, Scoring et Visualisation des résultats

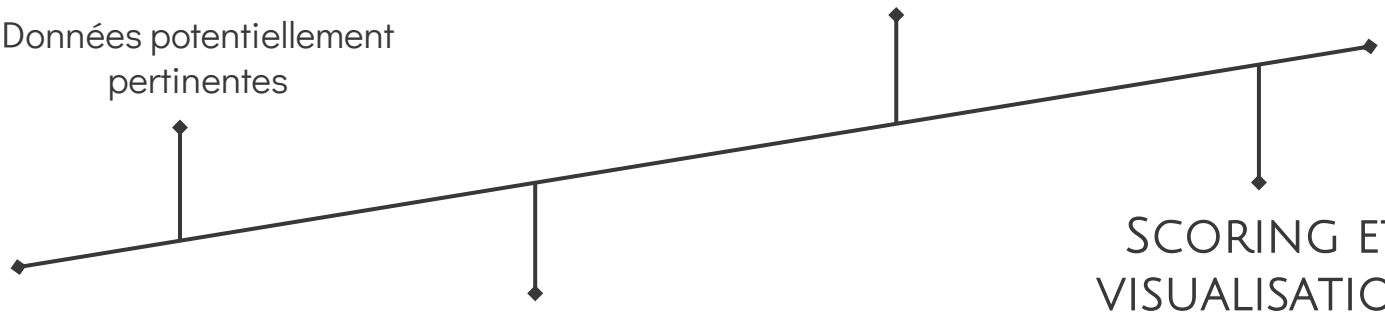
Stratégie

Pré-selection des données

Données potentiellement pertinentes

Sélection définitive

Choix des indicateurs et Pays finaux pour l'analyse



Analyse de la qualité des données

Taux de remplissage des données pré-sélectionnées

Scoring et visualisation

Analyse métier

Environnement de développement

Anaconda

Installation d'Anaconda: plateforme de disbution python la plus populaire

environement virtuel

Mise en place d'un envirennement virtuel dédié au projet

Installation des paquets

Installation des paquets nécessaires (numpy, pandas,matplotlib, seaborn) avec la commande **pip install**

Pré-sélection des données

Pré-sélection des données

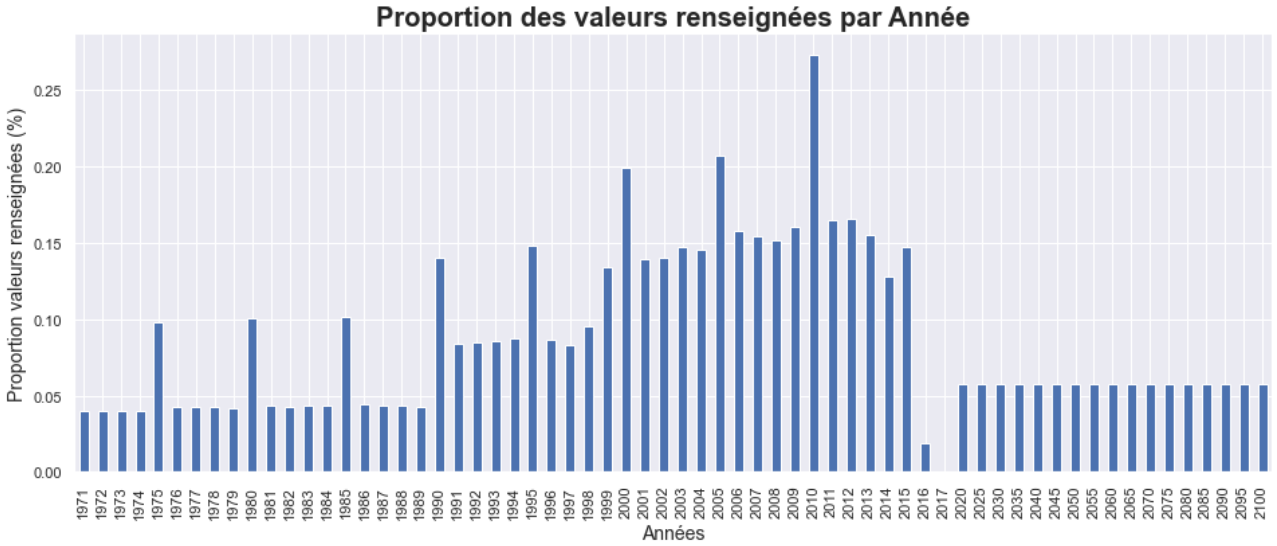
Rappels

- Métier: formation en ligne
- Cible: lycée, université
- Expansion à l'international

Fichiers importants

- EdStatsData.csv
- EdStatsSeries.csv
- EdStatsCountry.csv

Période pertinente



- Les données sont plus renseignées pour 2000-2015
- Ignorer les données 1970-1999

Topic pertinents

Topic	Description	Nb Indic.
Attaintment	Niveau d'instruction	733
Economic Policy & Debt: National accounts: US at current prices: Aggregate indicators	Pouvoir d'achat	3
Infrastructure: Communications	Utilisation d'internet	2
Health: Population: Dynamics	Croissance de la population	1
Health: Population: Structure	Structure de la population	13
Population	Structure population / niveau scolaire	213
Secondary	Niveau lycée	256
Tertiary	Niveau université	158
	TOTAL	**1379**

Répartition du dataset

- 7 Régions
- 242 Pays
- 1379 Indicateurs
- 8 Sujets (Topic)
- Annees 2000-2100

Historique

2000-2015

Projection

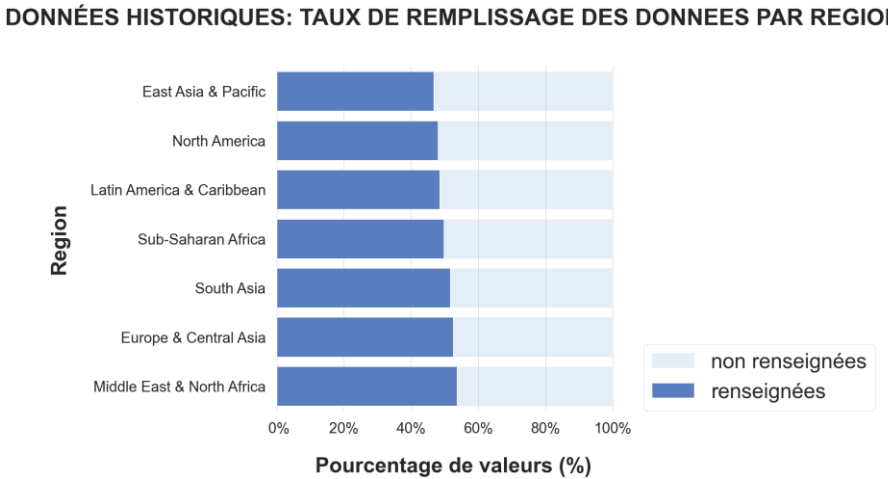
2020-2100

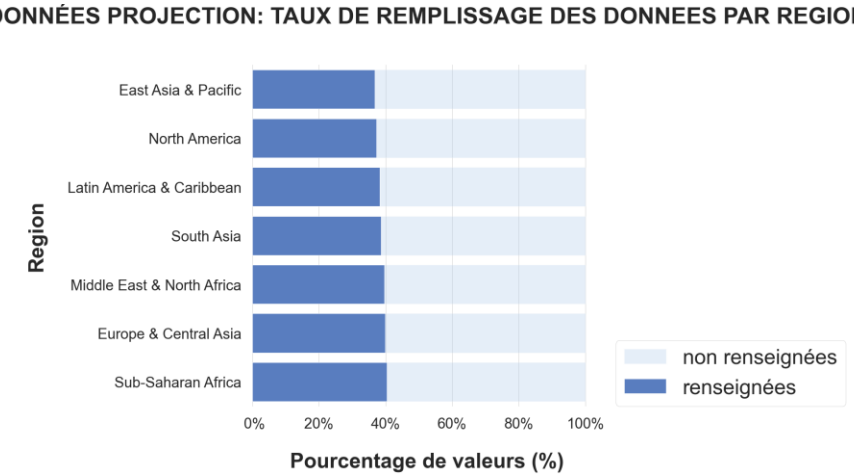
Données regroupées en 2 bases

- Les valeurs des indicateurs
- Par pays et région
- Sur les périodes

Analyse de la qualité des données présélectionnées

TAUX DE REMPLISSAGE PAR RÉGION

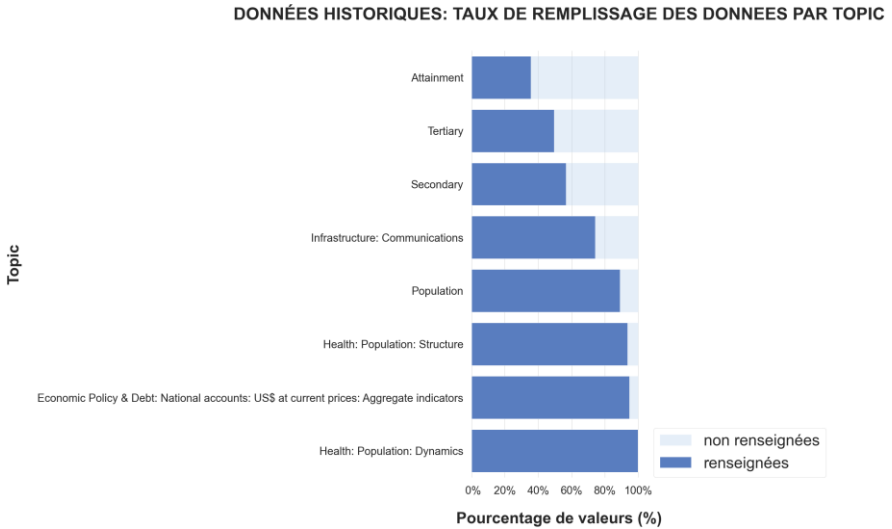


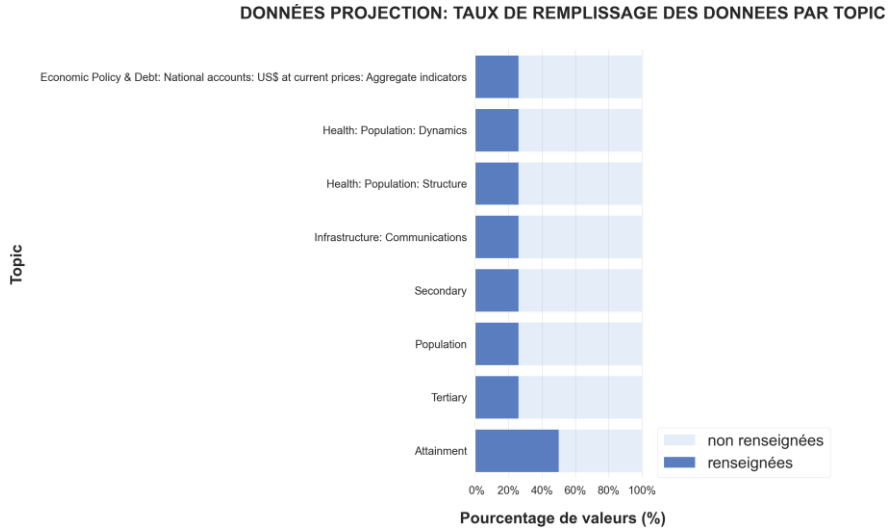


Données Historiques: taux de remplissage/région < 55%

Données de projection: taux de remplissage/région <= 40%

TAUX DE REMPLISSAGE PAR TOPIC



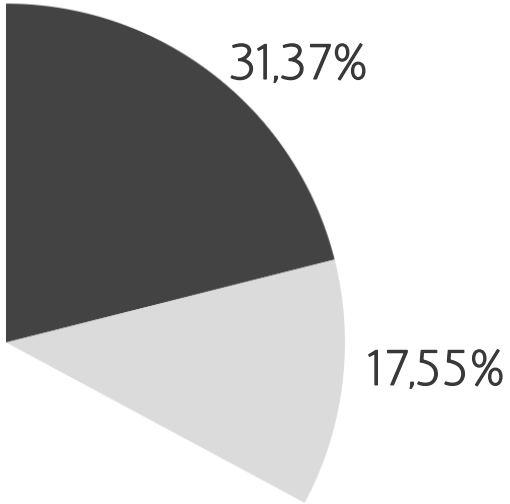


- Taux de remplissage > 50% pour la plupart des topics pré-sélectionnés
- Secondary et tertiary remplis (pourtant c'est la cible)

- Le topic Attaintment est apparement plus liés aux projections
- Les autres topics sont très peu remplis

Taux de remplissage global

- Historique
  2000-2015
- Projection
  2020-2100



Remarque

Jeu de données très peu rempli dans l'ensemble

Aucune ligne dupliquée dans le jeu de donnée

Sélection définitive des indicateurs et pays

INDICATEURS DE DESCRIPTION

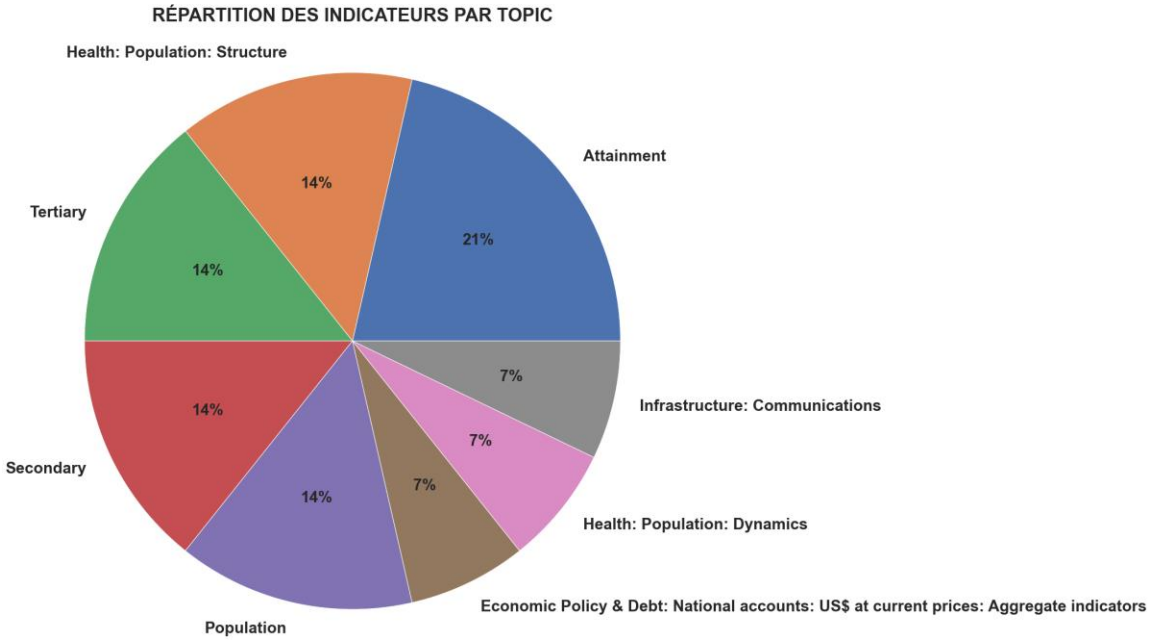
CODE INDICATEUR	TOPIC	SIGNIFICATION
NY.GDP.PCAP.CD	Economic Policy & Debt:…..	GDP per capita (current US)
IT.NET.USER.P2	Infrastructure: Communications	Internet users (per 100 people)
SP.POP.GROW	Health: Population: Dynamics	Population growth (annual %)
SP.POP.TOTL	Health: Population: Structure	Population, total
SP.POP.1564.TO.ZS	Health: Population: Structure	Population, ages 15-64 (% of total)
SP.SEC.TOTL.IN	Population	Population of the official age for secondary education, both sexes (number)
SP.TER.TOTL.IN	Population	Population of the official age for tertiary education, both sexes (number)
SE.SEC.ENRL	Secondary	Enrolment in secondary education, both sexes (number)
SE.SEC.ENRR	Secondary	Gross enrolment ratio, secondary, both sexes (%)
SE.TER.ENRL	Tertiary	Enrolment in tertiary education, both sexes (number)
SE.TER.ENRR	Tertiary	Gross enrolment ratio, tertiary, both sexes (%)

Indicateurs de projection

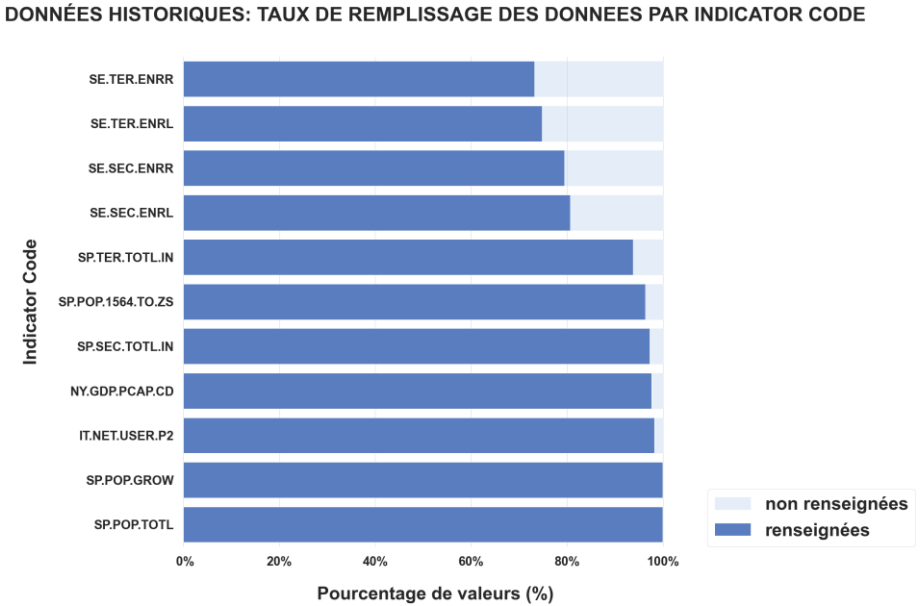
CODE INDICATEUR	TOPIC	SIGNIFICATION
PRJ.ATT.15UP.2.MF	Attaintment	Wittgenstein Projection: Percentage of the population age 15+ by highest level of educational attainment. Lower Secondary. Total
PRJ.ATT.15UP.3.MF	Attaintment	Wittgenstein Projection: Percentage of the population age 15+ by highest level of educational attainment. Upper Secondary. Total
PRJ.ATT.15UP.4.MF	Attaintment	Wittgenstein Projection: Percentage of the population age 15+ by highest level of educational attainment. Post Secondary. Total

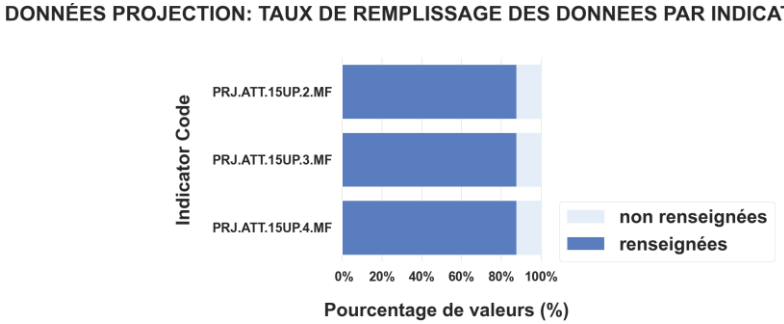
Répartion des indicateurs

- Internet
- Niveau lycée
- Niveau université
- Démographie
- Economie
- Projection



TAUX DE REMPLISSAGE DES INDICATEURS DÉFINITIFS





SELECTION DES PAYS

Suppression de tous les pays ayant moins de **30%** de taux de remplissage

Scoring et visualisation

Scoring

Objectif: attribuer un score à chaque Pays

Construction d'une dataframe dans laquelle les codes des indicateurs définitifs sont désormais en colonnes

Country Code, Country Name, Region, Year, list_Indicators

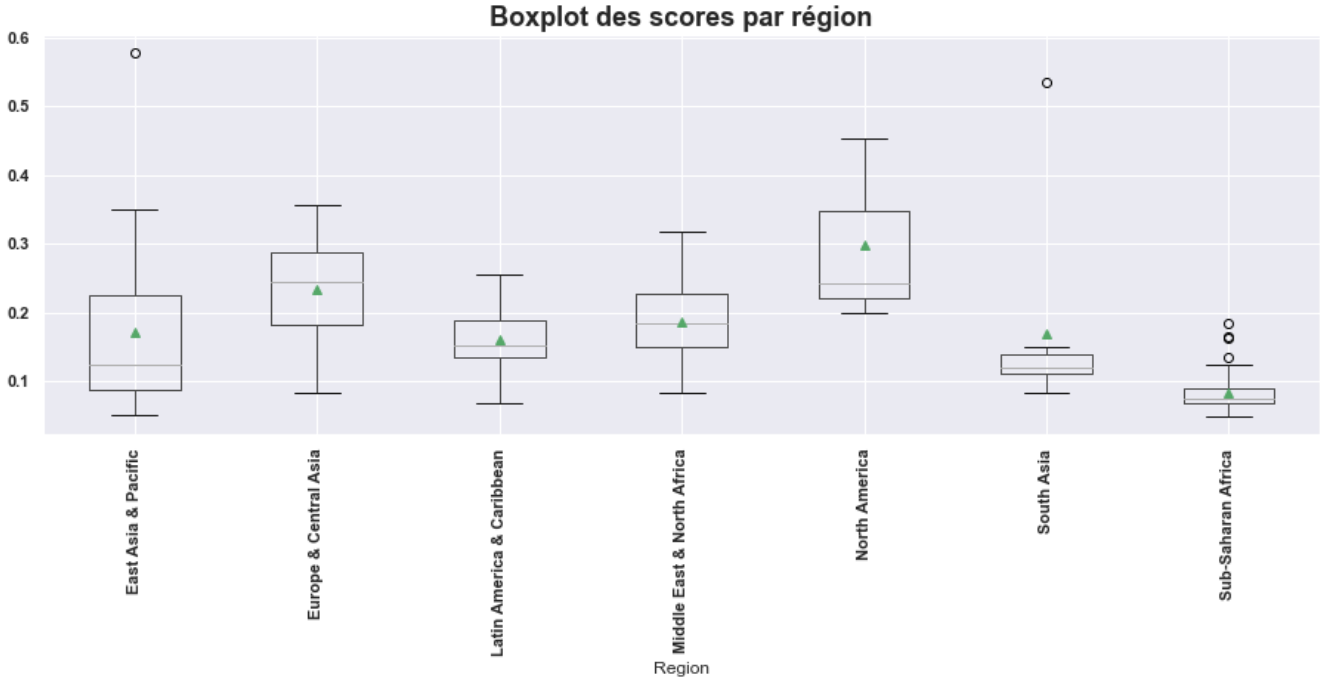
Normalisation min-max des valeurs de chaque indicateur (pour tous les pays) par années

Ajout d'une colonne «Score» qui contient la somme des valeurs de chaque ligne multiplié par un vecteur contenant les poids des indicateurs (1/11 pour tous les 11 indicateurs)

Interprétation du score

Plus le score d'un pays est élevé plus son potentiel de clients pour l'entreprise est grand

LES SCORES PAR RÉGION



POTENTIEL

Les régions North America et Europe & Central Asia semblent se démarquer des autres régions

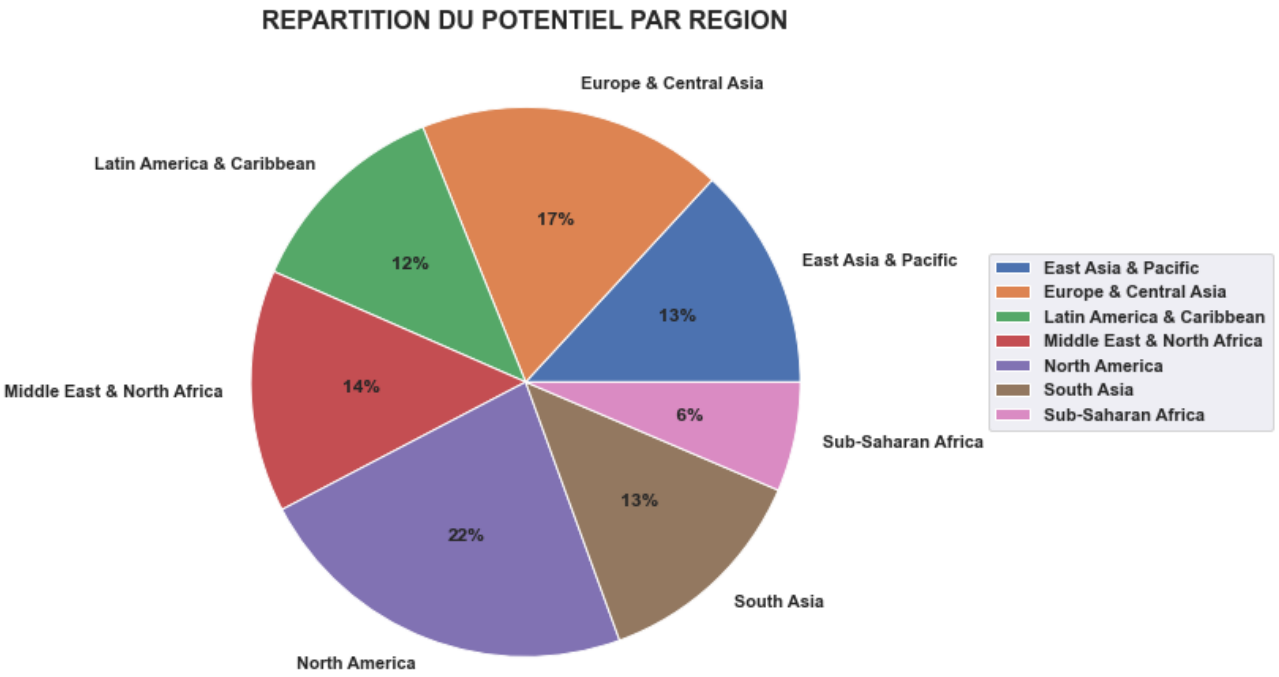
OUTLIERS

On observe des outliers (pays avec un score plus élevé par rapport aux autres pays de la région) pour les régions East Asia & Pacific, South Asia et Sub-Saharan Africa

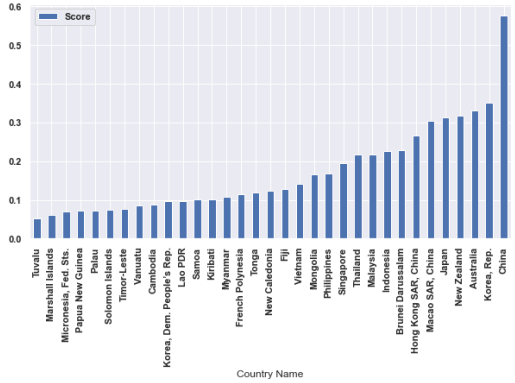
Potentiel par région (Score)

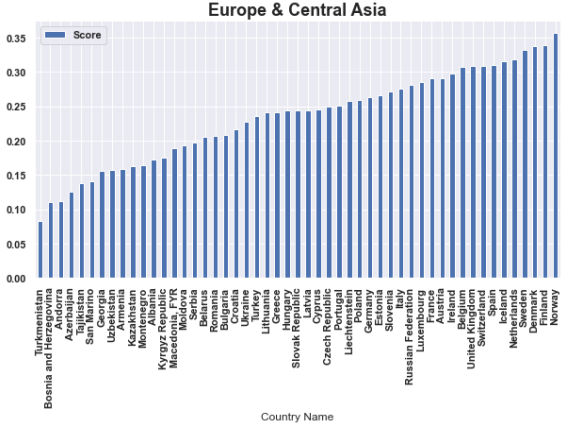
Potentiel

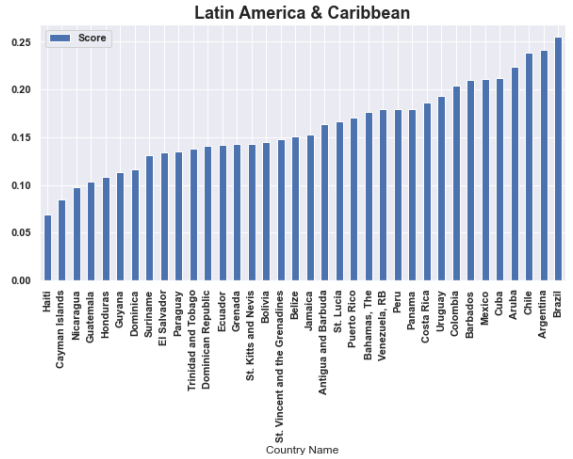
L'hypotèse émise avec le boxplot du score par région sur la démarquation des régions North America et Europe & entral Asia se confirme

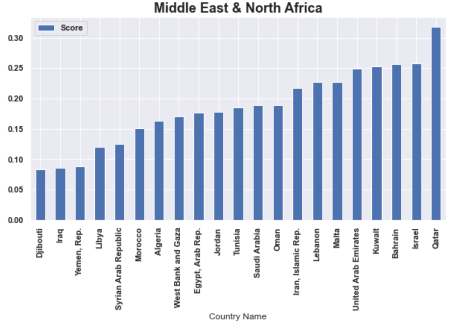


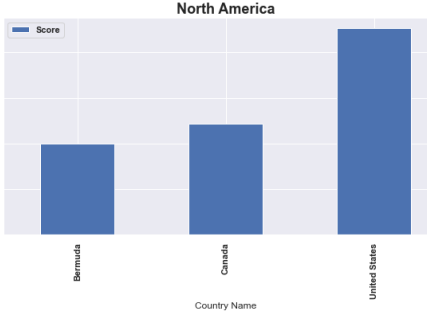
SCORES PAR PAYS

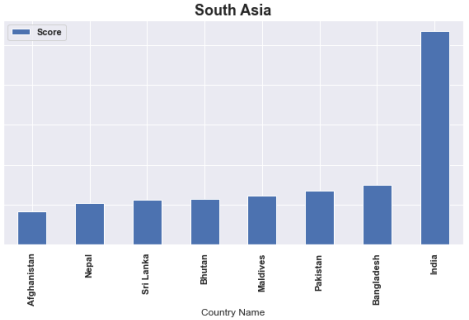


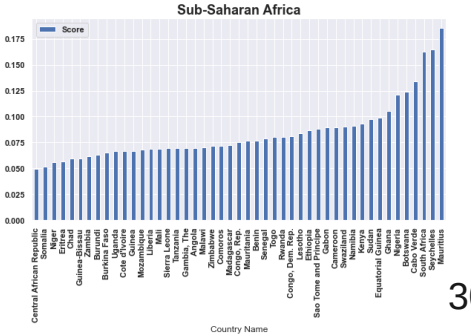




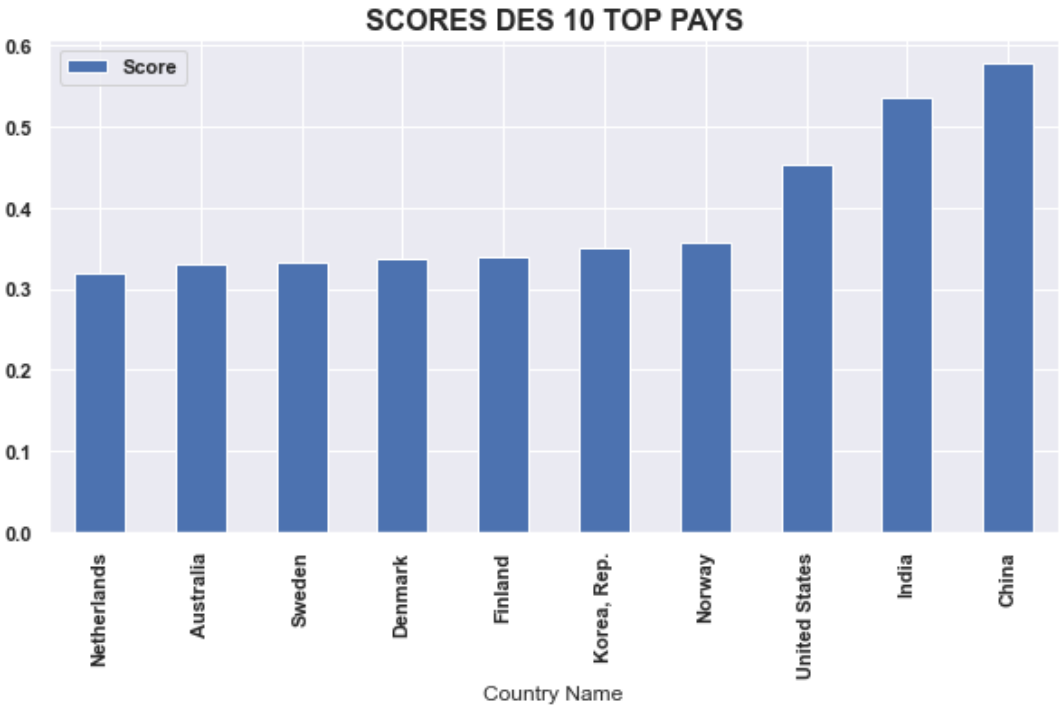


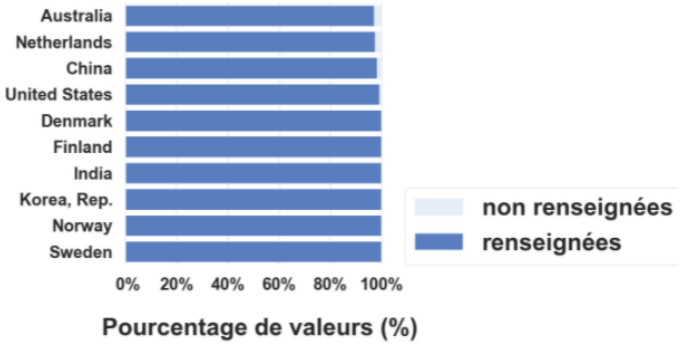






LE TOP 10 DES PAYS

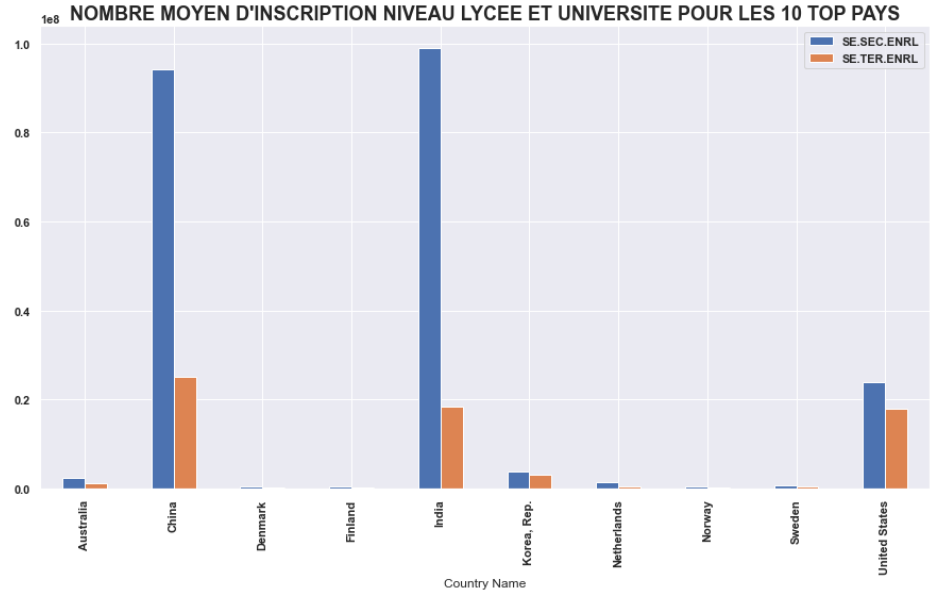


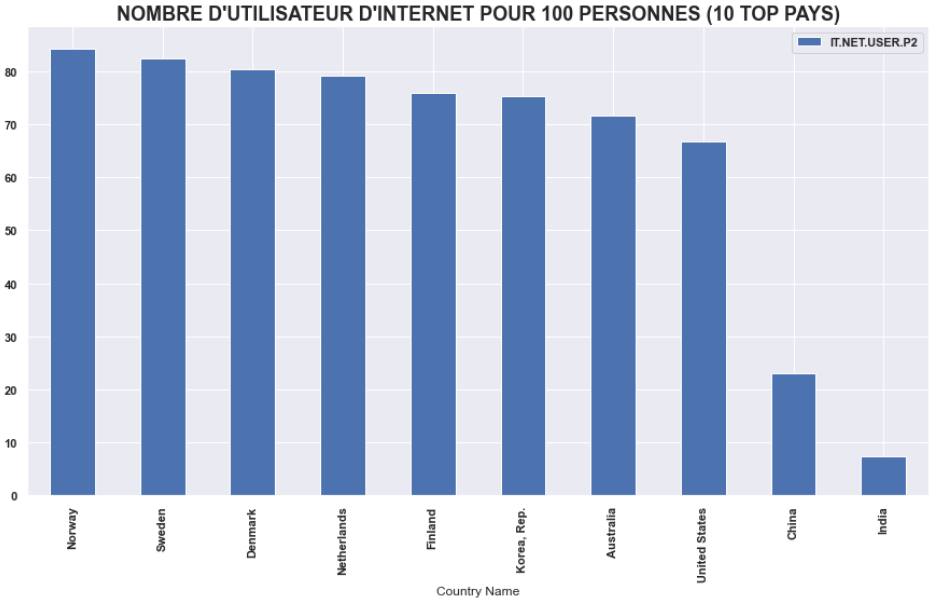


OBSERVATIONS

- La **Chine, l'Inde et les États-Unis** sont en tête du classement
- Les donnée sont très bien remplies pour ces pays (ce qui crédibilise les scores obtenus)

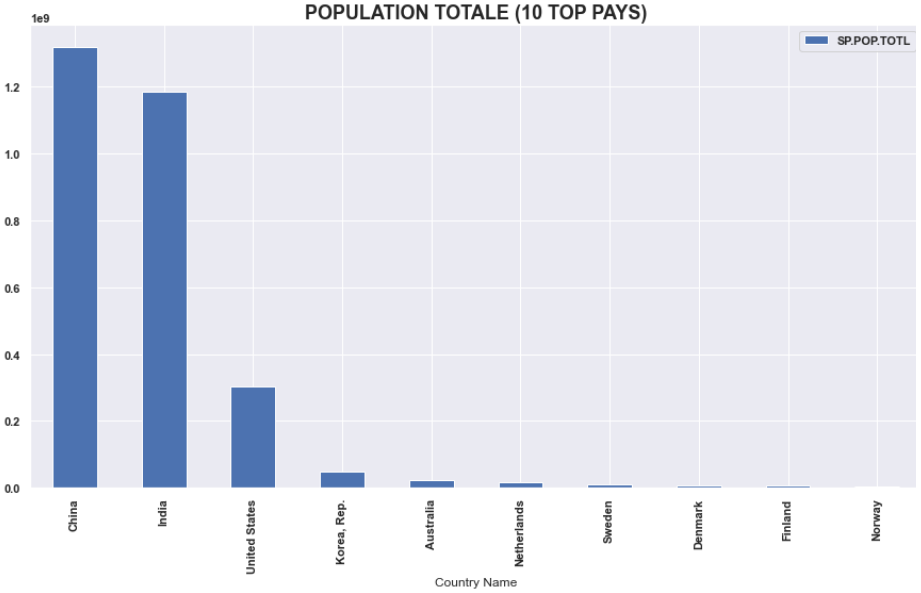
LYCÉE, UNIVERSITÉ, INTERNET

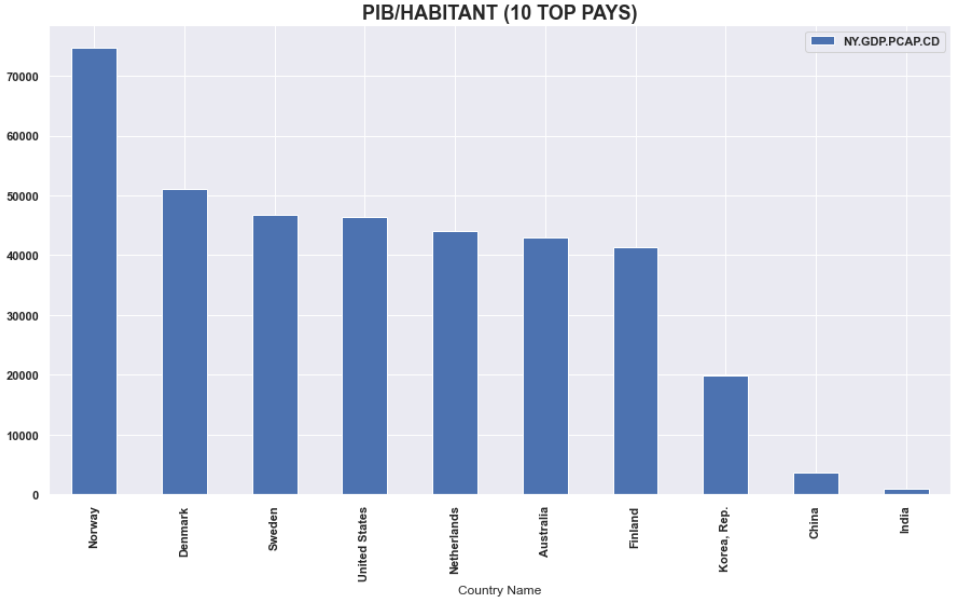




- La Chine et l'Inde ont un plus grand nombre de lycéens et d'étudiant
- Cependant ces 2 pays ont moins d'utilisateurs d'internet

POPULATION TOTALE, PIB/HABITANT





- La Chine et l'Inde ont une population totale très élevée
- Cependant la Chine et l'Inde ont les plus bas PIB/habitant

Les scores élevés pour la Chine et l'Inde semblent être beaucoup plus liés aux indicateurs relatifs à la démographie

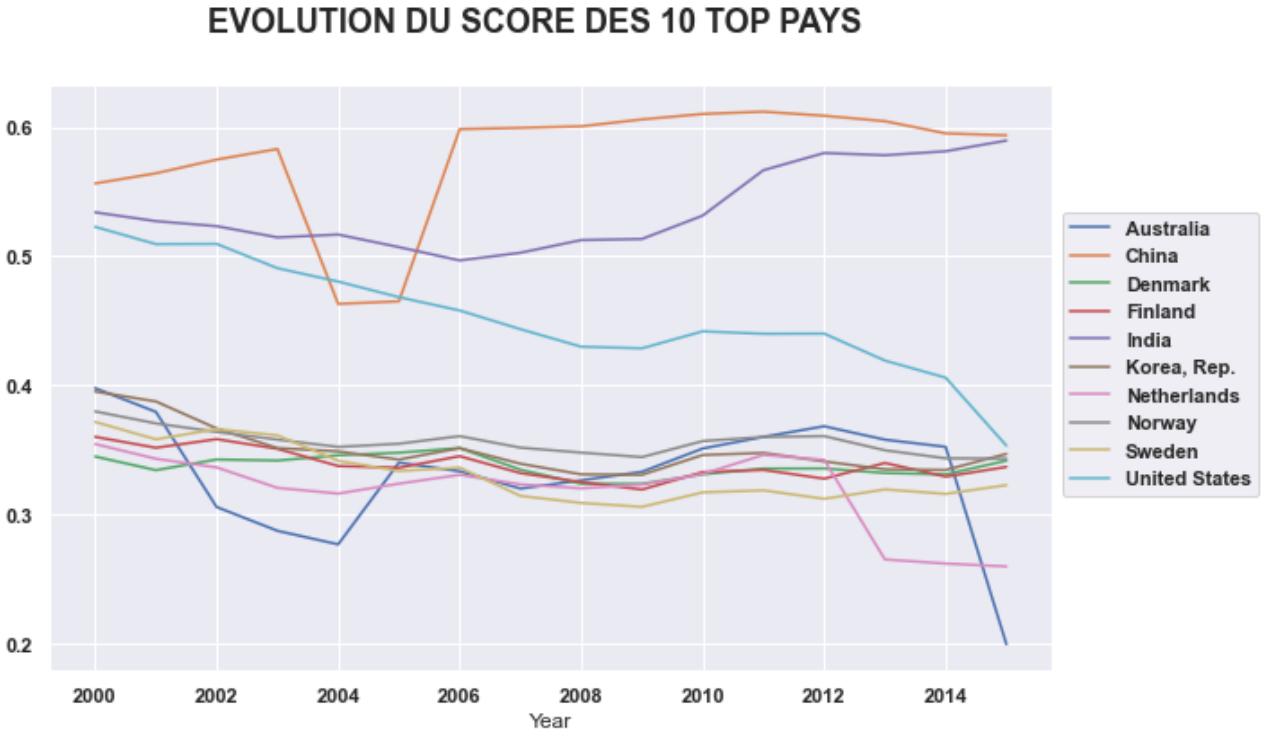
Evolution du potentiel

Observations

L'**Inde** semble être le pays ayant l'évolution la plus prometteuse

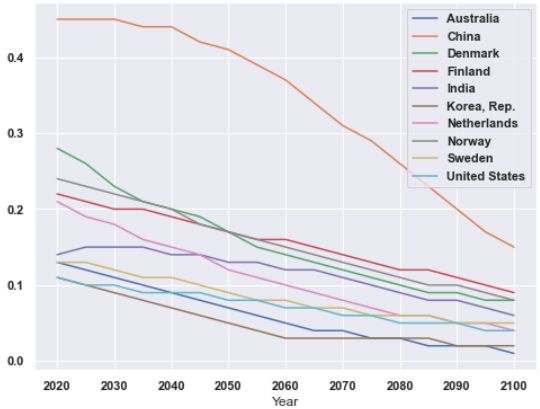
Tendance générale plutôt stable

A l'exception des Etats-Unis

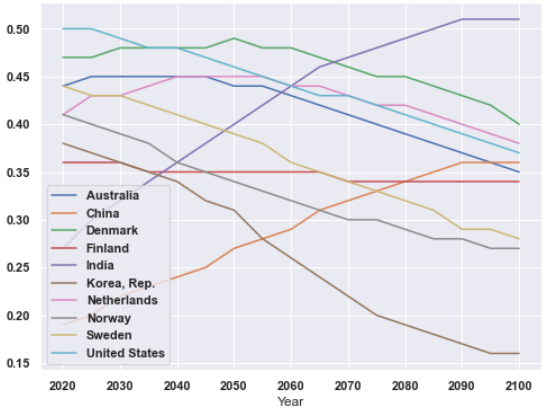


EVOLUTION INDICATEURS DE PROJECTION

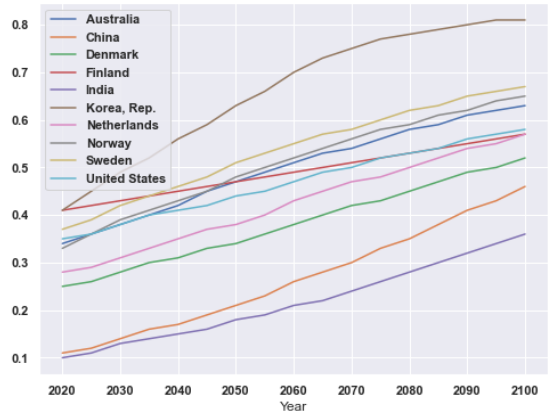
PROJECTION: EVOLUTION DU NIVEAU D'ETUDE (PRJ.ATT.15UP.2.MF)



PROJECTION: EVOLUTION DU NIVEAU D'ETUDE (PRJ.ATT.15UP.3.MF)



PROJECTION: EVOLUTION DU NIVEAU D'ETUDE (PRJ.ATT.15UP.4.MF)



Pour tous les 10 top pays, la tendance de l'évolution de la population de +15 ans ayant atteint le niveau "Post Secondary" sera croissante

03

Conclusion et perspectives

Suggestions de réponses aux questions métier et perspectives d'amélioration

Suggestions

Pays à fort potentiel

On pourrait sélectionner les pays avec les scores les plus élevés (Chine, Inde, Etats-Unis, Norvège, Corée, ...)

Evolution de ce potentiel

La tendance du score au fil des années est plutôt stable pour la plupart des top pays excepté pour l'Inde qui semblent avoir une tendance d'évolution croissante

Pays où il faudrait opérer en priorité

Si l'on s'arrête à ce que nous avons proposé, l'entreprise devrait opérer en priorité dans le pays avec les scores les plus élevés.
Par exemple: China, India, United States

PERSPECTIVES

- Sélection plus fine des indicateurs définitifs
- Attribution de poids différents aux indicateurs en fonction de leur niveau de pertinence d'un point de vue métier (par un expert du métier)

MERCI

Des questions ?